

The International Hapmap Project



Frank Dudbridge

MRC Biostatistics Unit, Cambridge, UK

What is the Hapmap?

- The Human Genome Project gave the “average” DNA sequence of a human
- This helps us find out how a human develops and works
- Does not show us the DNA differences between different humans
- These differences explain
 - Why people look and behave differently
 - History of populations and ethnic groupings
 - Why some people are more susceptible to certain diseases
- **The Hapmap is a catalogue of common DNA variation across the whole genome**

Who is the Hapmap?

- An international consortium of genotyping, statistics, sample collection and ethics groups
 - Canada, China, Japan, Nigeria, UK, US
- Total of 24 research groups
- Analysis by Oxford University, Wellcome Trust Centre for Human Genetics Oxford, Broad Institute, John Hopkins School of Medicine
- Data co-ordination by Cold Spring Harbor Laboratory
- Launched in 2002; 1st phase completed 2005

- **Key publications**
 - **Nature 426:789-796 (2003)**
 - **Nature 437:1299-1320 (2005)**
 - **<http://www.hapmap.org>**

The raw data

- Genotypes on SNPs in four populations
- 30 trios (two parents and an adult child) from Yoruba, Nigeria
 - Africans are less recent common ancestry than other continents
 - More genetic diversity
 - Less linkage disequilibrium
 - Yorubans are possibly the founding population for non-Africans
- 30 trios from Utah with European ancestry (from CEPH families)
- 45 unrelated Japanese from Tokyo
- 45 unrelated Chinese from Beijing
 - Japanese and Chinese turned out to be extremely similar
- Phase 1: one SNP every 5 kb, achieved 1,007,329 SNPs
- Phase 2: aims to genotype an additional 4.6M SNPs
 - Most of the common variation in Yorubans

Applications of the Hapmap

- Selecting SNPs for disease association studies
 - Tag SNPs
 - Genomewide SNP panels
- Selecting SNPs for admixture mapping
 - SNPs with different allele frequencies across populations
- Estimation of fine scale recombination rates
 - Estimated from an assumed coalescent model
 - Can estimate the rate without directly observing recombination
- Insights into natural selection
 - Long haplotypes as candidates for natural selection
 - Allele frequencies as conserved non-coding regions indicate purifying selection
- Developing methods for genetic association studies
 - Access to “real” data to create simulations for testing new methods

Haplotype tagging

- Suppose we have a gene of interest and want to see whether it contains a disease causing variant
- We will genotype SNPs in the gene and compare allele frequencies between disease cases and healthy controls
- Which SNPs should we genotype?
 - Expensive to resequence gene and genotype every variant
 - Better to find SNPs from the Hapmap – which covers most common variation
 - SNPs are correlated, so some SNPs can be predicted by the others
 - Pick a few SNPs which contain most of the diversity in the gene
 - These are called **tag SNPs**
- Tag SNPs are a cost effective way of scanning many genes
 - Even the whole genome

Browsing the Hapmap

- “Browse project data” from Hapmap home page
- First select a genome region

Search

[Help links:](#) - LD - - tagSNPs - - Phased Haplotype - - Genotype data - - Frequency data - - Symbols and colours used -

Landmark or Region :
chr21

Reports & Analysis :
Annotate LD Plot

Data Source
HapMap Data Rel#19/phaseII Oct05, on NCBI B34 assembly, dbSNP b124

- Then choose which features to display

Tracks

Overview All on All off

dbSNP SNPs/500Kb gt'd SNPs/500Kb Ideogram SNP cov/500Kb

Genes/500Kb Heteroz/500Kb NT contigs

Analysis All on All off

plugin:LD Plot plugin:Phased Haplotype Display plugin:tag SNP Picker

DNA All on All off

3-frame translation (forward) Contigs

3-frame translation (reverse) DNA/GC Content

Genes All on All off

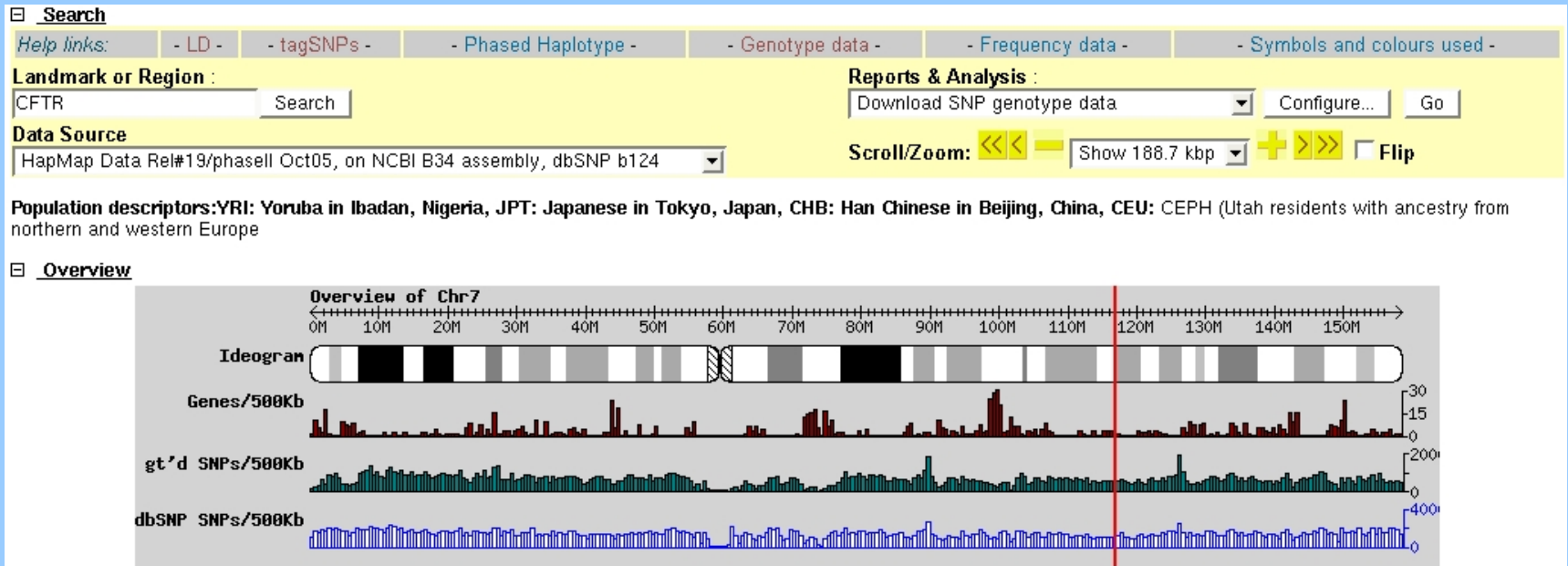
Entrez genes RefSeq mRNAs

Variation All on All off

dbSNP SNPs Heterozygosity/5Kb Recombination rate (cM/Mb) SNP coverage/5Kb

Genotyped SNPs Recombination hotspots Sequence Tagged Sites

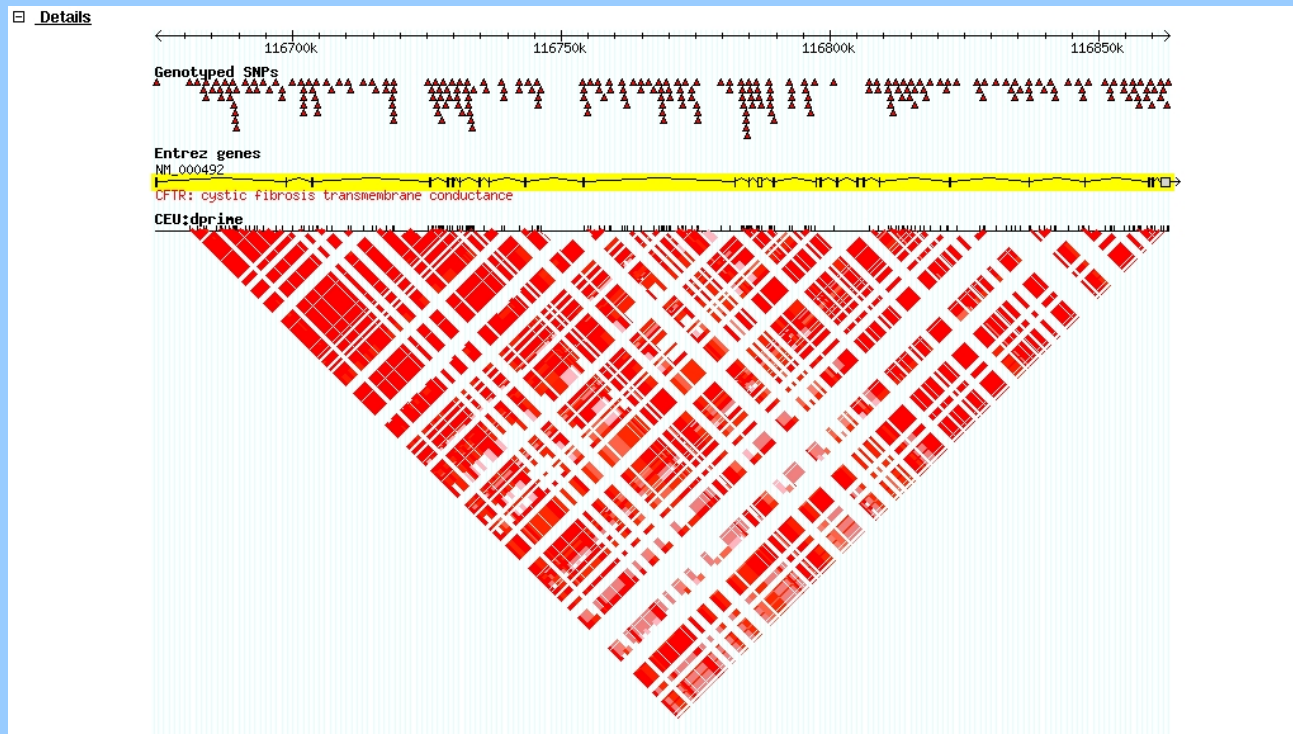
Chromosome overview



View the gene density, genotyped SNP density, dbSNP density etc

View LD pattern

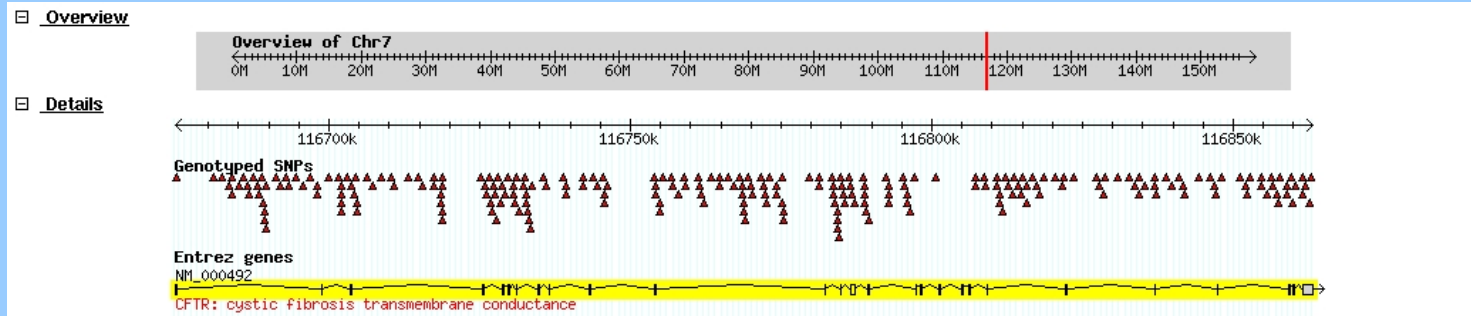
- Select “plugin: LD plot” and “genotyped SNPs”



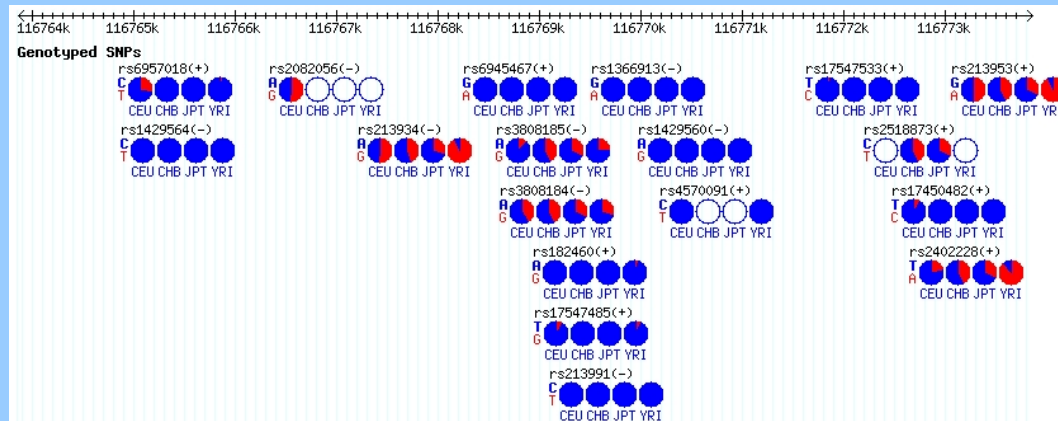
See how SNPs seem to group into blocks of high LD

We don't have to genotype all the SNPs in each block

View genotyped SNPs



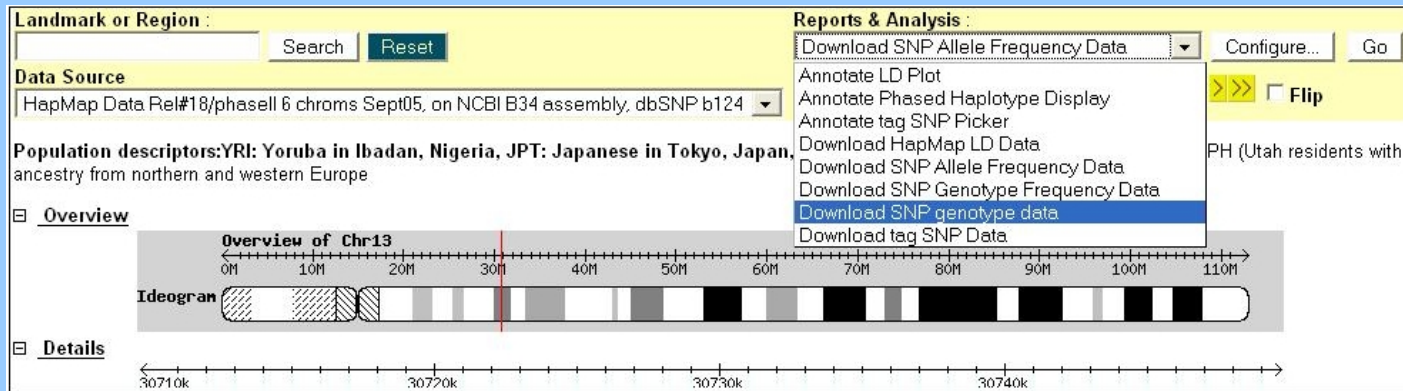
- At low resolution, see the density of genotyped SNPs



- At high resolution 10kb, see the allele frequencies in the four populations

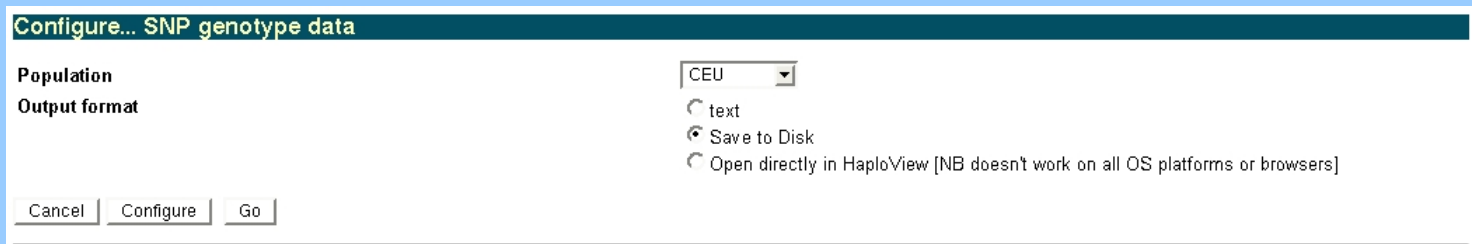
Download genotype data

- Download from the project data browser



The screenshot shows the HaploView interface. At the top, there is a search bar for 'Landmark or Region' with a 'Search' button and a 'Reset' button. Below this is the 'Data Source' section, which is set to 'HapMap Data Rel#18/phasell 6 chroms Sept05, on NCBI B34 assembly, dbSNP b124'. The 'Population descriptors' section lists 'YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, ancestry from northern and western Europe'. The 'Overview' section shows a chromosome ideogram for 'Overview of Chr13' with a scale from 0M to 110M. The 'Details' section shows a zoomed-in view of a region from 30710k to 30740k. The 'Reports & Analysis' dropdown menu is open, showing options: 'Download SNP Allele Frequency Data', 'Annotate LD Plot', 'Annotate Phased Haplotype Display', 'Annotate tag SNP Picker', 'Download HapMap LD Data', 'Download SNP Allele Frequency Data', 'Download SNP Genotype Frequency Data', 'Download SNP genotype data' (highlighted), and 'Download tag SNP Data'. There are also 'Configure...' and 'Go' buttons, and a 'Flip' checkbox.

- Select population and save to disk



The screenshot shows the 'Configure... SNP genotype data' dialog box. It has a title bar with the text 'Configure... SNP genotype data'. Inside, there is a 'Population' dropdown menu set to 'CEU'. Below it is the 'Output format' section with three radio button options: 'text', 'Save to Disk' (which is selected), and 'Open directly in HaploView [NB doesn't work on all OS platforms or browsers]'. At the bottom, there are three buttons: 'Cancel', 'Configure', and 'Go'.

- Saved in a format that can be read by Haploview
 - But not many other programs

Download using HapMart

- HapMart is graphical tool for exporting genotype and other data from Hapmap

bio::mart

new START FILTER OUTPUT export

count help

DATASET 1

FILTERS

MINOR ALLELE FREQUENCY [>=] 0.1

Monomorphic SNPs Only Excluded

SNPs found in 3'UTR Only Excluded

Limit to SNPs with these refIDs

REGION

Chromosome Chr18

From position

To position

GENE FILTERS

ENCODE REGION ENm010: 7:26730760..27230760

Summary

start

- Schema: rel20_NCBI_Build35
- Dataset: HapMap Population-CEPH(Utah)
- residents with Northern and Western European Ancestry

3719872 Entries Total

filter

- None

output

Not yet initialized

bio::mart

new START FILTER OUTPUT export

count help

Summary

Select the Attribute Page

GENOTYPE

Genotypes Details

SNP details

chromosome position

marker id alleles

reference allele genotyping center

genotyping platform

Genotype

CEU

Select the output format:

Text, fixed width Text, tab separated

Text, comma separated MS Excel

HTML

File compression:

None gzip (.gz)

Enter a name for this result set:

Name:

Enter a value to open results in new window (REQUIRES POP-UP UNBLOCKING), or to provide a name for file download.

Summary

start

- Schema: rel20_NCBI_Build35
- Dataset: HapMap Population-CEPH(Utah)
- residents with Northern and Western European Ancestry

53413 Entries pass Filters

filter

- Minor Allele Frequency [>=]: 0.1
- Monomorphic SNPs: Excluded
- Chromosome: Chr18

output

- GENOTYPE

53413 Results in Output

- Saved in a more easy to read format
 - But not directly usable by many programs

Haploview

- Haploview is a stand-alone program that is developed by HapMap partners
 - Not officially part of HapMap, but well integrated with it
 - Can read genotype dumps from HapMap
 - Linkage disequilibrium plots
 - Haplotype frequencies and blocks
 - Marker quality scores
 - Selection of tag SNPs
 - Association tests
-
- Not the fastest program around, nor including all the best methods
 - But a good starting point for analysing SNPs
 - As the authors admit
 - Bioinformatics 21(2):263-5

Haploview LD plot and haplotypes

- Display D' or r^2 values



- Display haplotypes in blocks and LD between blocks



Choose tag SNPs with Haploview

- Can force a SNP to be included or excluded
- Can select which SNPs to be tagged
- Show which tag SNPs tag which other SNPs

Haploview 3.2 -- sample.ped

File Display Analysis Help

LD Plot Haplotypes Check Markers Tagger Association

Configuration Results

#	Name	Position	Force include	Force Excl.	Capture this Allele
1	HGR1118a,1	274044	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
2	HGR1119a,1	274541	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
3	HGR1143a,1	286593	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
4	HGR1144a,1	287261	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	HGR1169a,2	299755	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
6	HGR1218a,2	324341	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
7	HGR1219a,2	324379	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
8	HGR1286a,1	358048	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
9	TSC01017	366811	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
10	HGR1373a,1	395079	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
11	HGR1371a,1	396353	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
12	HGR1369a,2	397334	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
13	HGR1369a,1	397381	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
14	HGR1367a,1	398352	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
15	HGR2008a,2	411823	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
16	HGR2008a,1	411873	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
17	HGR2010a,3	412456	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
18	HGR2011b,1	413233	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
19	HGR2016a,1	415579	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
20	HGR2020a,1	417617	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

pairwise tagging only r² threshold 0.8
 aggressive tagging: use 2-marker haplotypes LOD threshold for multi-marker tests 3.0
 aggressive tagging: use 2- and 3-marker haplotypes

Run Tagger Reset Table

Haploview 3.2 -- sample.ped

File Display Analysis Help

LD Plot Haplotypes Check Markers Tagger Association

Configuration Results

Tests

Allele	Test	r ²
IGR1218a,2	IGR1118a,1	0.952
IGR1369a,1	IGR1119a,1	0.949
IGR2008a,1	IGR1143a,1	1.0
TSC0101718	IGR1144a,1	0.954
IGR2020a,1	IGR1169a,2	0.894
IGR2008a,2	IGR1218a,2	1.0
IGR2011b,1	IGR1219a,2	0.908
	IGR1286a,1	0.898
	TSC0101738	1.0
	IGR1373a,1	0.954
	IGR1371a,1	0.952
	IGR1369a,2	1.0
	IGR1369a,1	1.0
	IGR1367a,1	0.907
	IGR2008a,2	1.0
	IGR2008a,1	1.0
	IGR2010a,3	0.814
	IGR2011b,1	1.0
	IGR2016a,1	0.853
	IGR2020a,1	1.0

Alleles captured by Current Selection

IGR1118a,1
IGR1119a,1
IGR1143a,1
IGR1144a,1
IGR1169a,2
IGR1218a,2
IGR1219a,2
IGR1286a,1
IGR1369a,1
IGR1367a,1

Captured 20 alleles with mean r² of 0.952
 Captured 100 percent of alleles with r² > 0.8
 Using 7 SNPs in 7 tests.

Dump Tests File

ENCODE regions

- Ten regions selected for Encyclopedia Of DNA Elements
- More complete discovery of common and rare SNPs
- Used to compare the Hapmap data to a more complete database
 - Assessment of completeness of the Hapmap
- Each region 500kb in length, resequenced in 48 subjects

- Density of SNPs is 10-fold higher in ENCODE regions
 - 1 SNP per 279 bp
 - Hapmap may only have 10% of all SNPs
- Most SNPs are rare: 46% with allele freq<5%
- But most heterozygous sites within individuals are due to common variation
- Uniform distribution of allele frequency in the main HapMap
 - Due to ascertainment bias in choosing common SNPs

- HapMap is sufficiently complete for non-African populations
- Should be complete for Africans after phase 2

Limitations of HapMap

- Population specific
 - Conclusions about the pattern of LD are specific to the four populations
 - Particularly relevant to African populations
 - Selection of tag SNPs and genomewide panel is population specific
 - However, similar conclusions between the four populations in HapMap
- Common variation
 - Conclusions about allele frequencies, natural selection biased by ascertaining common variation
 - Less of a concern in ENCODE regions
 - Less of a problem in phase 2
- Focus on SNPs
 - Insertion/deletion, copy number polymorphism may have relevance to common disease
- Questionable power of tag SNPs for association studies
 - Recent paper by Terwilliger & Hiekkalinna, Eur J Hum Genet 14:426-37

The future

- Phase 2 will be a more complete catalogue
- Genomewide association scans
 - Test loci across whole genome for disease association
 - Sufficiently dense map to detect LD with all disease loci
 - Minimal sufficient set of SNPs can be determined
- Population transferability and extension to other populations
- Development of better methods for genomewide analysis

The end